2020

# Classification Models of Idiopathic Pulmonary Fibrosis Patients

Mohammed Alqawba
*Old Dominion University*, malqa008@odu.edu

Luis R. Rodriguez

Norou Diawara
*Old Dominion University*, ndiawara@odu.edu

Rebecca T. Beuschel

Maryann Kaler

*See next page for additional authors*



## Original Publication Citation

Alqawba, M., Rodriguez, L., Diawara, N., Beuschel, R., & Kaler, M. (2020). Classification models of idiopathic pulmonary fibrosis patients. *International Journal of Respiratory and Pulmonary Medicine, 7*(1), 10 pp., Article 131. https://doi.org/10.23937/2378-3516/1410131

ISSN: 2378-3516

Alqawba et al. Int J Respir Pulm Med 2020, 7:131
DOI: 10.23937/2378-3516/1410131
Volume 7 | Issue 1
Open Access

International Journal of
Respiratory and Pulmonary Medicine

RESEARCH ARTICLE

# Classification Models of Idiopathic Pulmonary Fibrosis Patients

***Mohammed Alqawba[1], Luis R Rodriguez[2], Norou Diawara[1*], Rebecca T Beuschel[2], Maryann Kaler[3], Amisha V Barochia[3], Stewart J Levine[3], Steven D Nathan[4] and Geraldine Grant[2]***

[1]*Department of Mathematics & Statistics, Old Dominion University, USA*

[2]*Department of Biology, George Mason University, USA*

[3]*Laboratory of Asthma and Lung Inflammation, Pulmonary Branch, Division of Intramural Research, National Heart, Lung and Blood Institute, National Institutes of Health, Bethesda, USA*

[4]*Advanced Lung Disease and Transplant Program, Inova Heart and Vascular Institute, Falls Church, USA*

***Corresponding author:** *Norou Diawara, Department of Mathematics & Statistics, Old Dominion University, Norfolk, VA, USA*

## Abstract

Idiopathic pulmonary fibrosis (IPF) is a chronic and fatal interstitial lung disease with no current cure. Progression of IPF is difficult to predict as the clinical course can be highly variable and range from a rapidly deteriorating state to a relatively stable state, or may be characterized by a slow progressive decline. Therefore, the need for an accurate diagnosis and improved tools for monitoring and managing IPF is of paramount importance, all for understanding the mitochondrial structure and the function played in the IPF. Mitochondrial DNA copy number (MtDCN) has been correlated with mortality in IPF patients and is a source of potentially clinically relevant information. We investigated the effects of various expiratory variables on MtDCN via multiple linear regression models. The models and their theoretical framework are presented under a descriptive and then analytic approach to investigate the complex and impact causes of IPF. Generalized linear model (GLM) based boosting is fitted before and after imputing the missing data. The Bayesian Hierarchical logistic models with categorical response variables that were created using carefully chosen cut-off points to classify the patients. This research provides an opportunity for novel patient surveillances.

## Keywords

Idiopathic pulmonary fibrosis, Modelling, Prediction, Classification

## Introduction

Idiopathic pulmonary fibrosis (IPF) is a chronic and fatal interstitial lung disease with no current cure. The overall incidence of this disease is estimated at 6.8 to 17.4 cases per 100,000 people, with a 10-fold increased incidence in individuals over the age of 65. Progression of IPF is difficult to predict as the clinical course can be highly variable and range from a rapidly deteriorating state to a relatively stable state, or may be characterized by a slow progressive decline [1]. Until recently there was no pharmacological therapy for IPF. However, in October 2014 the first two FDA approved therapies (nintedanib and pirfenidone) were introduced into the market. These therapies have variable efficacy and may be associated with a variety of side effects. Therefore, the need for an accurate diagnosis and improved tools for monitoring and managing IPF is of paramount importance. The search for blood biomarkers that can aid in the diagnosis and monitoring of disease progression remains elusive. Mitochondrial DNA copy number (MtDCN) has been correlated with mortality in IPF patients and is a source of potentially clinically relevant information for physicians [2-6]. To date, there is little literature that explores the correlations between MtDCN and IPF factors and disease progression. We propose several models that investigate the effects of MtDCN on patients with IPF. The objective of this study is to build predictive models that correlate biomarkers derived from the peripheral blood of IPF patients, with clinical lung function data. This paper

**Citation:** Alqawba M, Rodriguez LR, Diawara N, Beuschel RT, Kaler M, et al. (2020) Classification Models of Idiopathic Pulmonary Fibrosis Patients. Int J Respir Pulm Med 7:131. doi.org/10.23937/2378-3516/1410131

**Accepted:** March 14, 2020; **Published:** March 16, 2020

**Table 1:** List of clinical and molecular markers meausre collected from the patient cohort.

| **Summary of abbreviated data terms** | | |
|---|---|---|
| Clinical measure | Notation | Delta at 3, 6, 12 months |
| Predicted forced vital capacity (FVC%) | dp_d1 | dp_3, dp_6, dp_12 |
| Predicted forced expiratory volume (FEV1%) | d1p_d1 | d1p_3, d1p_6, d1p_12 |
| Diffusing capacity of the lung for carbon dioxide (DLCO%) | DLCO | DLCO_3, DLCO_6, DLCO_12 |
| Gender, age, lung physiology index (GAP) | GAP | |
| Body mass index (BMI) | BMI_d1 | BMI_3, BMI_6, BMI_12 |
| Six minute walk test (6MW) | w6MW_d1 | wp_3, wp_6, wp_12 |
| Mitochondrial DNA/genomic DNA ratio (Mt DCN) | Y1 | |

studies the differences between the values of MtDCN of normal subjects and IPF patients. We describe further factors that might impact the values of MtDCN in IPF patients.

We first investigated the effects of various expiratory variables on MtDCN via multiple linear regression models after performing variables selection. Data adjustments were made to satisfy the normality assumption of the ordinary regression model and imputation of the missing data was done using the Bayesian bootstrap method [7]. To validate the analysis, Generalized linear model (GLM) based boosting was fit before and after imputing the missing data. The results show consistency in estimating the effects of the chosen predictors. Next, to avoid the normality assumption, we fit Bayesian Hierarchical logistic models with categorical response variables that were created using carefully chosen cut-off points to classify the patients.

The remaining paper presents the models and their theoretical framework in construct and hypothesis building, under a descriptive and then analytic (epidemiology) approach to investigate the complex and impact causes of IPF.

## Background

The aims of this research is to understand how clinicians can internally compute and communicate quantitative variables-such as changes in mitochondrial DNA copy counts-and use such information to guide clinical management. The observation that mitochondrial DNA copy counts decrease with both age and disease is well validated [8-10]. However, the link between disease severity and changes in this marker have yet to be fully elucidated. By comparing this marker to clinical and demographic data in 30 IPF patients we expect to identify a small number of quantitative variables that can be expanded to a larger cohort and used to establish meaningful clinical indicators.

IPF is a highly heterogeneous disease that has a variable and unpredictable clinical course in individual patients. Clinical factors such as age, sex, lung function, and smoking history have an association with the progression and severity of disease. Yet, conceptual frameworks for thinking about such processes, and our understanding of their medical implementation, remain in their infancy. This study includes 67 participants over the age of 50 of which 31 have been clinically diagnosed with IPF. Each patient donated 10 ml of whole blood from which Peripheral blood mononuclear cells (PBMCs) were isolated.

The IPF patients who participated were referred to the Advanced Lung Disease Program at Inova Fairfax and diagnosed with Idiopathic Pulmonary Fibrosis based on the current consensus guidelines [11]. The study was approved by the Inova Fairfax Hospital Internal Review Board (IRB #06.083) with appropriate written informed consent obtained for each patient. Normal participants included presented with no clinical history of IPF and a normal chest X-ray. All normal subjects provided informed consent to participate in protocol 15-H-0017, which was approved by the NHLBI IRB.

Mitochondrial DNA copy number was calculated following a previously published protocol [5]. We collated clinical data from multiple sources obtained over the course of 1-year prior to the blood sampling. Our clinical data includes standard demographic information such as: Age, sex, familial status, smoking status, and life time smoking volume (Table 1).

More attention will be given to the statistical models and their extensions. The power and flexibility to manipulate data (missing or incomplete) and accommodate the error components provide calibration to patients' classification and medical researchers in general will be provided next.

## Statistical Methods

Model specification and assumption can elucidate the data. Exploratory analysis is performed and covariates selected. Generalized linear models [12] with low Akaike information criteria (AIC) will be deemed satisfactory and the covariates used in the further analyses. Markov Chain Monte Carlo methods will be used to assess variational inference and follow them up with classification and predictive accuracy.

First, we may transform the data of the latent variables in our model to the real coordinate space and en-

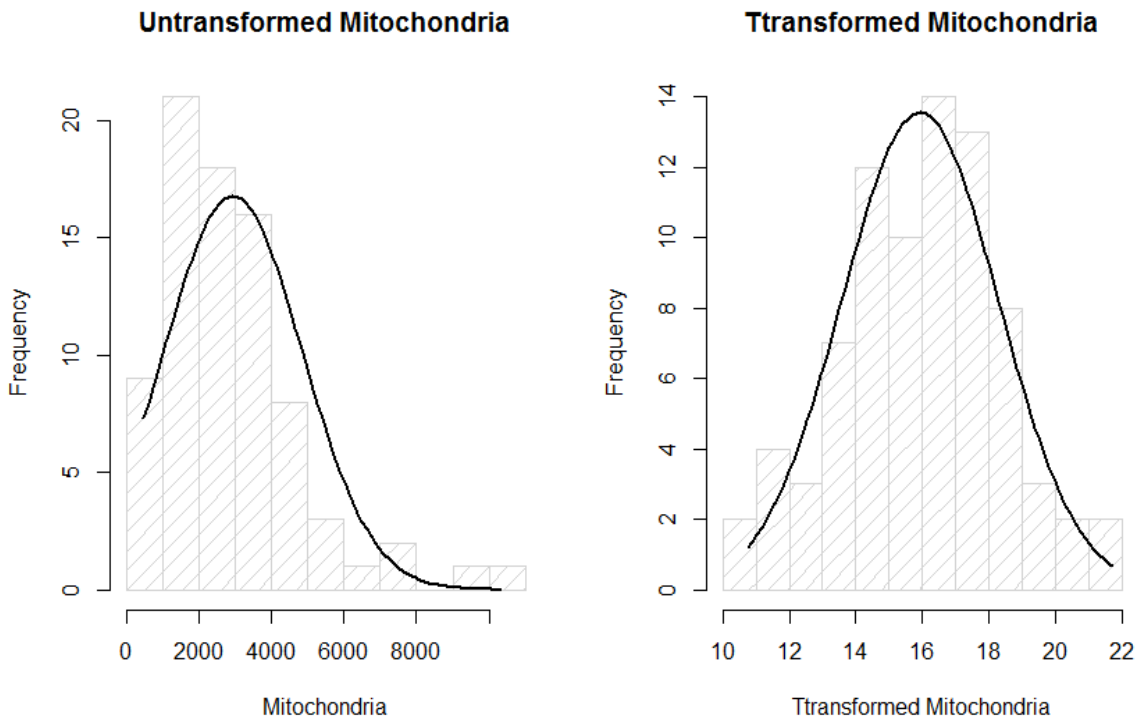


**Figure 1:** Normality plots of mitochondria.

sure normality. For example, the logarithm transforms a positively constrained variable, with a standard deviation, to the real line. Then, we plan to posit a Gaussian variational distribution. This induces a non-Gaussian approximation in the original variable space. Last, we combine automatic differentiation with stochastic optimization to maximize the variational objective. We begin by defining the class of models we support. Bayesian bootstrap imputation procedure will be used as it is an efficient method to handle missing data in a multilevel setting.

## Comparison between normal participants and IPF patients

We are interested in finding a predictive model of the response MtDCN, say *Z*. To fit linear regression model, the response has to be distributed as normal distribution. Running the Shaprio's Test of normality, with the null hypothesis as $H_0$: *Z* is normally distributed, we get $W = 0.88144$ and small p - value < 0.05. That suggests *Z* is not normally distributed. Therefore, Box-Cox transformation method was considered to obtain normally distributed variables that represent MtDCN (Figure 1).

The new variables are given as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

Where λ is estimated using MLE method that is of value 0.164576. Now, that the assumption of normality of the response *Z* after transformation denoted as *Y*, has met normality, we fit linear regression and ANOVA without violated the important normality assumption of the data.

For all patients, the box plots show that there is a huge variation among the oxygen users from their smoking status. As the effect of smoking as a risk factor or indicator in IPF is under debate, for the purpose of this investigation, we assumed that former smokers have greater severity in the MtDCN markers than non-smokers (Figure 2).

To have a better understanding of the estimates of the variances, and use the answers to guide better inferential questions, a table of analysis of variance formally called ANOVA is presented to summarize the data at first.

One-way ANOVA test (F-test) of *Y* (MtDCN) based on whether the subject is sick or not, we find the following:

model is: $Y_{ij} = \mu + \tau_i + e_{ij}$,

where $\mu$ is the overall mean of $Y_1$,

$\tau_i$ is the effect of group and

$e_{ij}$ is the error term for each observation

with $i = 0$ *or* $1$ and $j = 1,2,\ldots, 36$ for normal subject = 0

and $j = 1,2\ldots, 30$ for IPF patient = 1.

The outputs of the ANOVA table are given in Table 2. It shows that the two groups of study participants IPF (IPF patients and normal subjects) are significantly different, with a p-value of 0.0155.

The mean sum of square errors that accounts for the variance that is unexplained after accounting for the linear effect of the patients' group is of 5.83. The Table 2 result shows that the test distinguishes between the IPF disease and non-disease individuals

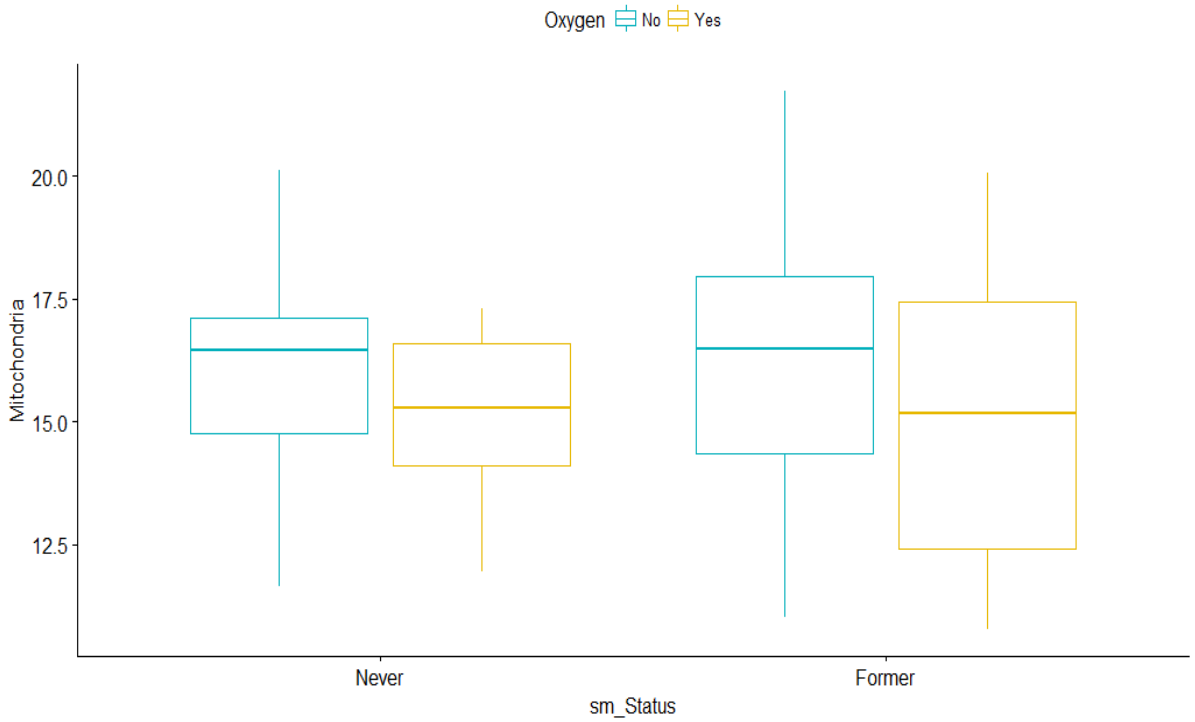


**Figure 2:** Bar graphs of Mitochondria *vs.* Smoking status and oxygen use.

**Table 2:** Analysis of variance table for Mitochondria *vs.* IPF patient groups.

| | df | SS | MS | F value | P-value |
|---|---|---|---|---|---|
| IPF Status | 1 | 36.0 | 36.0 | 6.179 | 0.0155 |
| Residuals | 65 | 378.7 | 5.83 | | |

**Table 3:** Regression line of Mt DNCN *vs.* 6MW.

| | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 12.909 | 1.298 | 1.298 | < 0.0001 |
| w6MW_d1 | 0.006 | 0.002 | 0.003 | 0.0214 |

**Table 4:** Individual linear regression models of MtDNC and IFP patient groups.

| | | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|---|
| Normal | Intercept | 11.4286 | 2.356 | 4.851 | < 0.0001 |
| | w6MW_d1 | 0.00929 | 0.0042 | 2.201 | 0.0346 |
| IPF patient | Intercept | 14.1985 | 1.8343 | 7.740 | < 0.0001 |
| | w6MW_d1 | 0.00203 | 0.0039 | 0.513 | 0.6120 |

and that the subject status of whether being sick or not is significant with p-value of the ANOVA equals 0.0155. This shows that there is significant difference between the two groups of IPF patients and normal subjects/participants. The means for each group turns out to be $\bar{Y}_1$ = 16.57038 and $\bar{Y}_2$ = 15.10032 are the means of MtDCN for normal and IPF patients, respectively. However, it is not clear that the biomarker variables will guide in the prediction of the disease status.

A formal straight line is then considered under the linear regression models of MtDCN against the predictors available for both normal and IPF groups. We found the following:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$$

where $Y$ is the transformed MtDCN, and $X_1$ is the 6MW (m) predictor. The estimates are reported in Table 3. The results indicate that the 6MW significantly affects MtDCN.

Since the two groups of participants (IPF patients and normal subjects are significantly different), we propose to compute their individual linear regression models. The outputs of the models are described in Table 4.

Here the dependent variable w6MW_d1 is more desirable in predicting the MtDCN for normal subjects than for IPF patients. Note that the variance for the w6MW_d1 is quite high for normal patients.

**Table 5:** Estimates parameters of simple regression of Mt DCN *vs*. DLCO.

| | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 13.60768 | 0.86115 | 15.802 | < 0.0001 |
| DLCO | 0.03590 | 0.01275 | 2.817 | 0.00642 |

**Table 6:** Individual linear regression models of Mt DCN and DLCO based on IFP patient groups.

| | | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|---|
| Normal | Intercept | 15.56318 | 2.12070 | 7.339 | < 0.0001 |
| | DLCO | 0.01235 | 0.02569 | 0.481 | 0.634 |
| IPF patient | Intercept | 12.64840 | 1.98104 | 6.385 | < 0.0001 |
| | DLCO | 0.05747 | 0.04485 | 1.281 | 0.21 |

**Table 7:** Multiple linear regression of Mt DCN vs selected predictors'outputs.

| | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 7.1030 | 4.10442 | 1.731 | 0.0954 |
| dp_12 | -0.2014 | 0.07636 | -2.637 | 0.0139 |
| dp_3 | 0.2534 | 0.1034 | 2.450 | 0.0213 |
| Sm Status | 2.0194 | 1.0475 | 1.928 | 0.0649 |
| DLCO | 0.0751 | 0.0548 | 1.371 | 0.1822 |

This is to point out that the reasonable accuracy from knowledge can lead to suspicious disturbance in the quest for model procedure, strategy and simplicity.

Another added variable DLCO may add better fit to the MtDCN response variable. It is tested next.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_2$$

where $Y$ is the transformed MtDCN, and $X_2$ is the DCLO predictor.

As it turns out, the estimated variance of the model has decreased, while model significance has been maintained (Table 5).

As earlier, a break down by participants is also suggested, and the results are described in Table 6. Surprisingly, the results indicate that the DLCO variable does not significantly affect MtDCN within each of the grouping. Also, other variables were not significant. This shows the complexity behind modelling such data phenomena, and a simple model may not be accurately describing the patients' information. We will consider more extensive modelling techniques.

## Multiple linear model approach

Since we are interested in MtDCN, and since systematic difference is neither obvious nor predictive, multiple linear model approach may help address discrepancies and variable selection.

We start the analysis of $Y$ (transformed MtDCN) being regressed on the full set of explanatory variables. Under stepwise regression method, the following variables were selected, and they are described by this model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4$$

where $X_1$ is the dp_12 e.g. *delta FVC%* at the 12th month,

$X_2$ is dp_3 e.g. *delta FVC%* at is the at the 3rd month,

$X_3$ is the sm_status

$X_4$ is the DLC

The output is displayed in the Table 7. Note that DLCO is not significant with p-value of 0.1822. Other statistics that are used as measures of goodness of fit are usually $\hat{\sigma}^2$ = 6.75 and *AIC* = 153.7.

The multiple linear model seems reasonable. However, as most patients do not continue smoking after disease diagnosis, it seems reasonable to consider smoking i.e. sm_status as a weight instead of a variable itself. Looking at an ANOVA test (t-test) of $Y_1$ based on smoking, we find the following:

model is: $Y_{ij} = \mu + \tau_i + e_{ij}$,

where $\mu$ is the overall mean of $Y_1$,

$\tau_i$ is the effect of smoking level and

$e_{ij}$ is the error term for each observation

with $i$ = 2 or 3 and $j = 1,2,\ldots, 11$ for never smoker = 2

and $j = 1,2,\ldots, 14$ for former smoker = 3.

The outputs of the ANOVA test are displayed in Table 8.

The Table 7 result showed that the sm_status is not significant with p-value of the ANOVA equals 0.0649, and is confirmed in Table 8. This evidences that there is no significant difference between the two levels of sm_status. The means for each group turns out to be $\bar{Y}_1$ = 3.89 and $\bar{Y}_2$ = 3.98 are the means of MtDCN for

**Table 8:** Analysis of variance table for Mitochondria *vs*. smoking status.

| | df | SS | MS | F value | P-value |
|---|---|---|---|---|---|
| Sm status | 1 | 0.0492 | 0.04921 | 1.413 | 0.247 |
| Residuals | 23 | 0.8012 | 0.03484 | | |

**Table 9:** Multiple linear regression of Mt DCN *vs*. dp_12 and dp 3' outputs.

| Variable | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|
| Intercept | 15.23549 | 0.51435 | 29.621 | < 0.0001 |
| dp_12 | -0.18907 | 0.07863 | -2.405 | 0.0231 |
| dp_3 | 0.21915 | 0.10482 | 2.091 | 0.0457 |

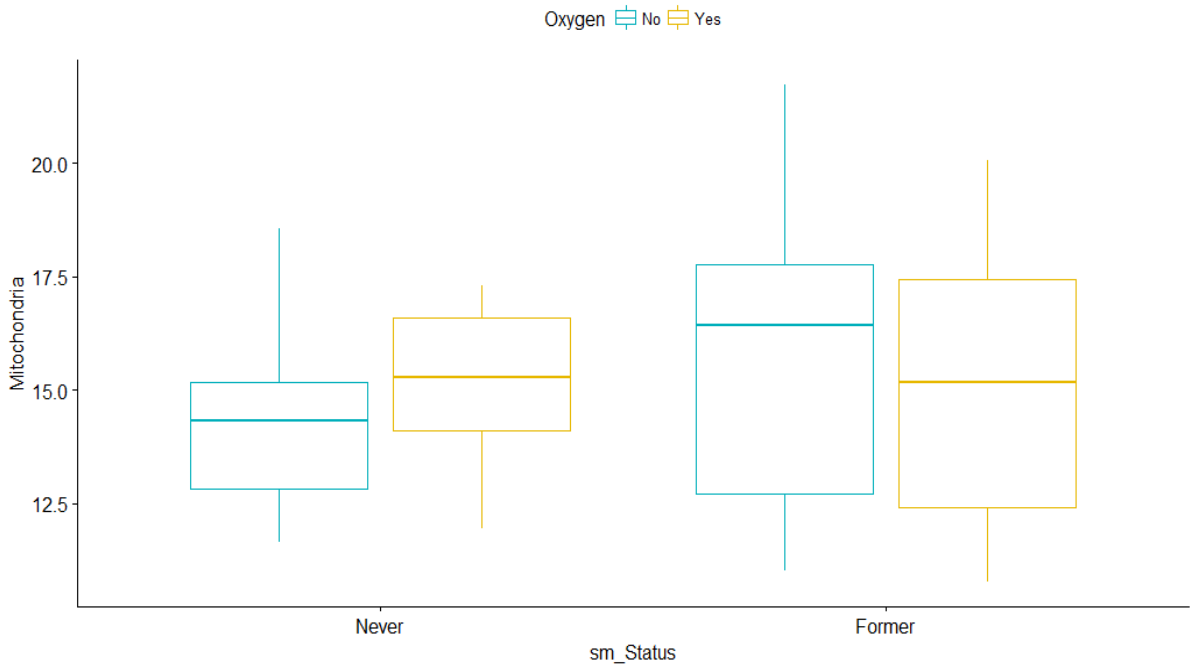


**Figure 3:** Bar graphs of Mitochondria *vs*. Smoking status and oxygen use for IPF patients only.

sm_status = 2 and sm_status = 3, respectively.

The plots in Figure 3 show the linear relationship between the response $Y_1$ and each one of the predictors to further understand the difference between the two smoking groups based on oxygen usage. It also shows that MtDCN is higher among oxygen users than among non-oxygen users for patients who never smoke. However, the MtDCN for oxygen users is lower for former smokers than for non-oxygen users. The problem is that a patient assigned to oxygen would be quite random, counterfactual sometimes. To determine causal effects for these patients is counterproductive, as the prediction is purely observational. Rather, we will try to capture indirect effects.

We then consider different predictors and try to understand their impacts. The effects of dp_12 and dp_3 on MtDCN are displayed in the next graphs, while keeping the smoking status in mind. It is observable that MtDCN values of non-smokers seem to be enclosed within the MtDCN values of former smokers when estimated with dp_3 and dp_12. This fact will be review as a doughnut or torus case in the later section, as it reveals a circular concentration. In the meantime, we explore further and add more predictors at hand. Dropping DLCO, we get the following model

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

where $X_1$ is the dp_12 e.g. *delta FVC%* at the 12th month and

$X_2$ dp_3 e.g. *delta FVC%* at is the at the 3rd month.

The output of the model is displayed in Table 9.

From plots in Figure 4, we can see that the never smoker group is surrounded by the former smoker group. In fact, when we model each group separately we get different results in term of significance.

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Similarly, we consider the variable oxygen, which takes the value 1 if a patient takes oxygen, and 0 otherwise. Although including the variable oxygen in the model does not lead to a good model, dividing the data into two groups based on the oxygen status shows

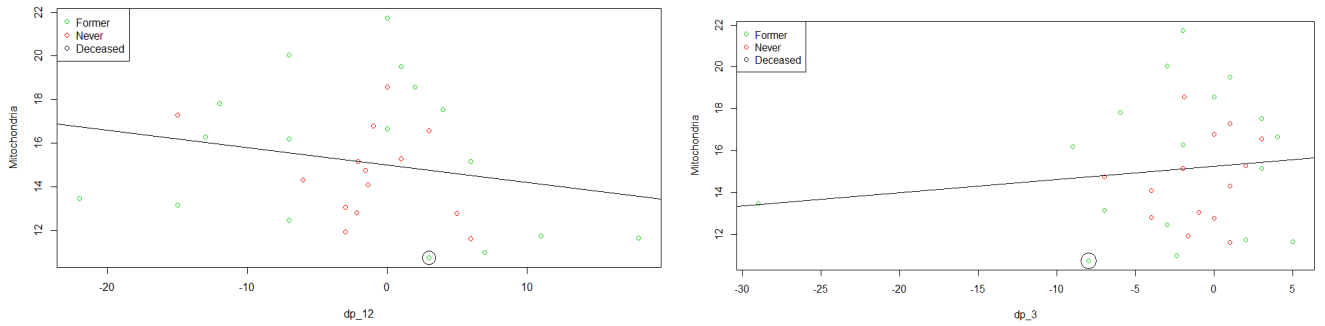


**Figure 4:** Plots of MtDCN *vs*. dp_12 and dp_3.

**Table 10:** Individual linear regression models of Mt DCN and DLCO based on selected predictors.

| | Variable | Estimate | Std. Error | t value | P-value |
|---|---|---|---|---|---|
| Sm status = Never | Intercept | 14.40782 | 0.59652 | 24.153 | < 0.0001 |
| | dp_12 | -0.15683 | 0.11116 | -1.411 | 0.186 |
| | dp_3 | 0.09695 | 0.20854 | 0.465 | 0.651 |
| Sm status = Former | Intercept | 16.0223 | 0.8359 | 19.167 | < 0.0001 |
| | dp_12 | -0.2381 | 0.1137 | -2.094 | 0.0549 |
| | dp_3 | 0.3000 | 0.1471 | 2.040 | 0.0607 |
| Oxygen = 0 | Intercept | 15.4821 | 0.7390 | 20.950 | < 0.0001 |
| | dp_12 | -0.2388 | 0.1523 | -1.569 | 0.136 |
| | dp_3 | 0.2415 | 0.1574 | 1.534 | 0.145 |
| Oxygen = 1 | Intercept | 14.96505 | 0.75444 | 19.836 | < 0.0001 |
| | dp_12 | -0.19056 | 0.09215 | -2.068 | 0.0686 |
| | dp_3 | 0.36663 | 0.21889 | 1.675 | 0.1283 |
| Pirfenidone = 1 | Intercept | 14.4578 | 0.7815 | 18.501 | < 0.0001 |
| | dp_12 | -0.2359 | 0.0986 | -2.392 | 0.0404 |
| | dp_3 | 0.2024 | 0.1869 | 1.083 | 0.3068 |
| Nintenadib = 2 | Intercept | 15.2714 | 0.6652 | 22.956 | < 0.0001 |
| | dp_12 | -0.1903 | 0.1131 | -1.682 | 0.113 |
| | dp_3 | 0.2278 | 0.1355 | 1.681 | 0.114 |

that predictors $dp_{12}$ and $dp_3$ significantly affect MtDCN among those who take oxygen, which is not true with those who do not take it. The model indicates that there are other confounding factors besides oxygen and the dp_3 and dp_12.

Finally, we fit the model for each group based on antifibrotic status (Pirfenidone and Nintenadib). The results are shown in Table 10. The model shows a significant effect of dp_12 on the MtDCN for Pirfenidone. However, the model does not dissuade from adding the dp_3, and the idea that there are disturbances and variational errors within the patients, and thus further analysis should be considered.

Plot in Figure 5 allows to assess how well the fitted multiple regression model describes the IPF prediction based on smoking and also based on oxygen use. Although the predictive model is capturing the MtDCN, the classification is not at all clear. Hence, we use the techniques of boosting and Bayesian models.

## Boosting approach

Parameter estimations comparison between standard GLM and GLM-based boosting before and after estimating the missing values using multiple imputations under Bayesian bootstrap (BB):

Consider the linear model

$$\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

where $Y_1$ is the transformed value of MtDCN,

$X_1$ is the dp_12 e.g. *delta FVC%* at the 12th month and

$X_2$ dp_3 e.g. *delta FVC%* at is the at the 3rd month.

Before handling the missing data, the parameter estimates are given in Table 11.

After handling the missing data using multiple imputations under Bayesian bootstrap, the parameter estimates are provided in Table 12.

From the results, the estimates from the GLM

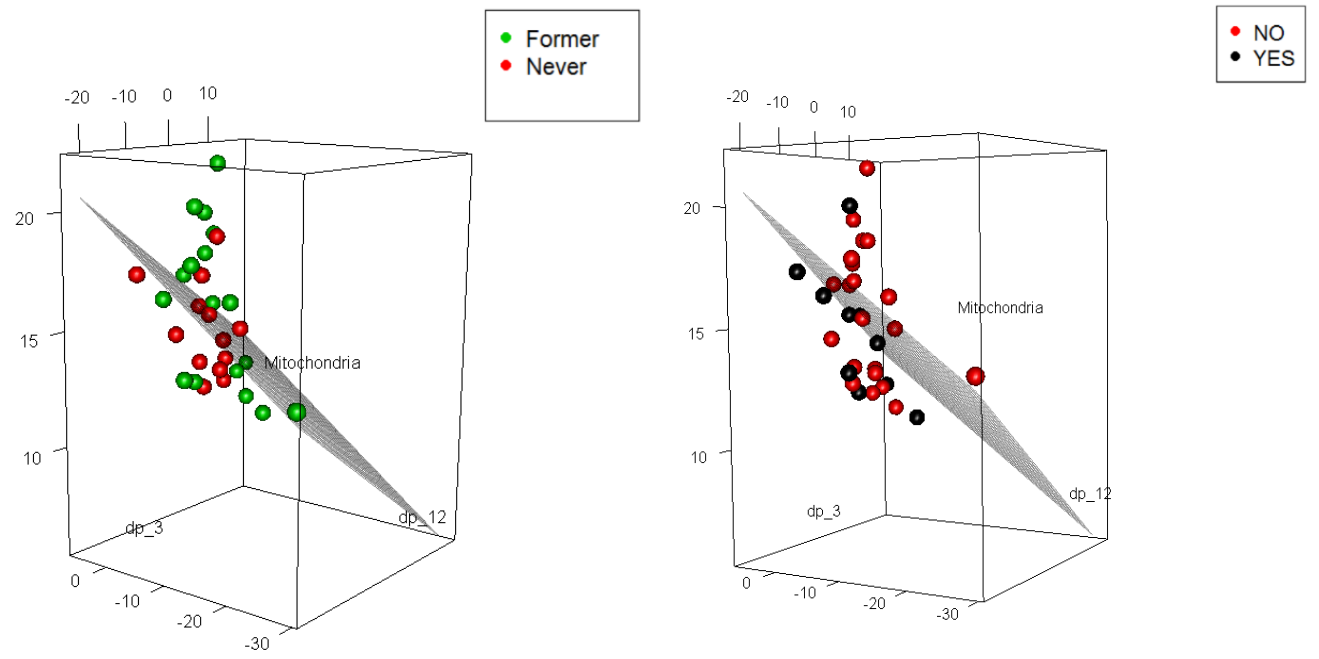


**Figure 5:** 3d plots of MtDCN with smoking (2 levels, but one patient was deceased) and oxygen use (2 levels).

**Table 11:** Estimates from two models.

| Parameter | Standard GLM | GLM-based boosting |
|---|---|---|
| $\hat{\beta}_0$ | 4.678712 | 4.67696129 |
| $\hat{\beta}_1$ | -0.020076 | -0.016893103 |
| $\hat{\beta}_2$ | 0.023374 | 0.019063909 |

**Table 12:** Estimates from two models with imputed data.

| Parameter | Standard GLM | GLM-based boosting |
|---|---|---|
| $\hat{\beta}_0$ | 4.656684 | 4.654382 |
| $\hat{\beta}_1$ | -0.01914114 | -0.01642997 |
| $\hat{\beta}_2$ | 0.0226934 | 0.01894534 |

based with missing data boosting algorithm are close to the estimates from the standard GLM, hence showing that our method is robust.

## Bayesian hierarchical logistic regression model with mixed effect

To obtain and challenge the formulation above and have a reduced form equation of the variables, we consider the Bayesian hierarchical regression model, from the conditional indirect effects. The GLM used is extended to include a Bayesian setting. We start by classifying the patients into groups determined by their MtDCN levels. Then, we build predictive model by asking if the patients can be classified as sick or very sick based on their dp_3 and dp_12, adding MtDCN levels, Bayesian.

Several works have proposed strategies for computations of under optimal designs [13,14], and stopping criteria in a missing data context. Examples are presented in both [15,16], Such technique will be applied to the features tracking and classification. That is, we will represent MtDCN by an indicator variable associated with DLCO that significantly affects the MtDCN values of both IFB and normal subjects.

Although we transformed the response variable (i.e. MtDCN) to satisfy the normality assumption of the linear model, the assumption is not quite satisfied (refer to the Q-Q plots). Thus, another technique in modeling the IFP of both patient and normal participants is considered to model MtDCN indirectly to avoid strong distributional assumption. Analysis on the threshold of the biomarkers is further performed on the IPF patients. The idea behind introducing an indicator, which classify the subjects into two groups, as it appears that a removal of DLCO as an effect on the MtDCN, may lead to overfitting, or confounding due to dp_12 or dp_3.

This indicator we defined here is as follow:

$$z = \begin{cases} 1, & if\ DLCO < median\,(DLCO), \\ 0, & otherwise, \end{cases}$$

The idea behind introducing this indicator, which classify the subjects into two groups, comes from the following plots where we can see when we plot MtDCN against DLCO, we can easily distinguish between the two groups of subjects. It appears that a decreased level of DLCO has an effect on the MtDCN.

Now we will fit a Bayesian Hierarchical logistic model with $Z$ as a response and MtDCN and DLCO or 6mw as

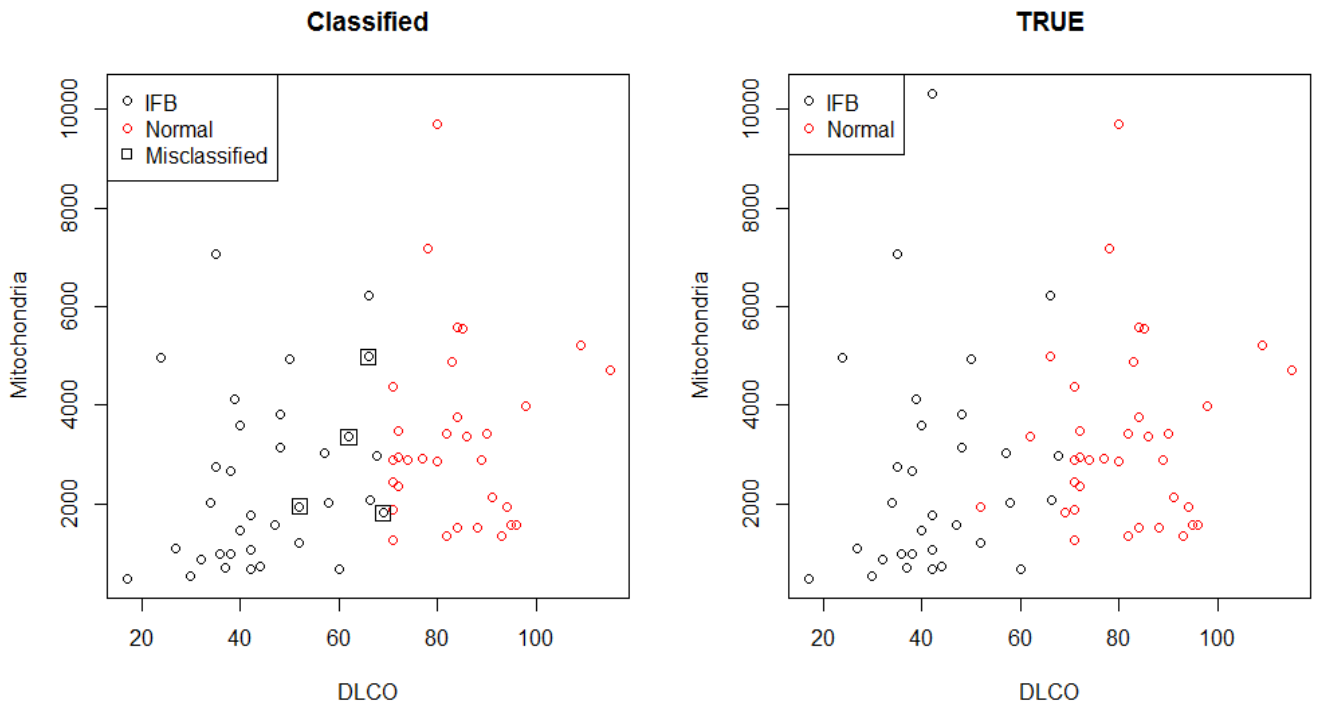


**Figure 6:** Classification comparison.

**Table 13:** Estimates of model classification parameters based on Mt DCN and DLCO.

| | Mean | Std. Error | Rhat |
|---|---|---|---|
| $\hat{\beta}_0$ | 1046.889 | 1546.947 | 1.027 |
| $\hat{\beta}_1$ | -0.001 | 0.014 | 1.015 |
| $\hat{\beta}_2$ | -15.433 | 22.412 | 1.027 |
| $\sigma^2_{sm_{status}}$ | 22.743 | 41.963 | 1.012 |

**Table 14:** Classification outcomes.

| | | Predicted | |
|---|---|---|---|
| | | 0 | 1 |
| True | 0 | 32 | 4 |
| | 1 | 0 | 31 |

predictors. The model is given by

$\Pr(Z_i = 1) = logit^{-1}(X_i'\boldsymbol{\beta} + \alpha_j)$, $i = 1,\ldots,80$, $j = 1,2$

where $\alpha_j^{sm_{status}} \sim N\left(0,\ \sigma^2_{sm_{status}}\right)$,

$\beta \sim N(0, \Sigma)$.

For the first model, X = (1, $x_1$ = Mitochondria, $x_2$ = 6MW) and $= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \sim N\left(\mathbf{0}, \Sigma = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}\right)$

The model expression in such that whether or not the participant has an overexpressed level of DLCO with covariates as MtDCN and 6mw. The model expression including the random component is useful as the clinical course among patients is unpredictable and with high variability (Table 13).

Classification table of the predicted normal and IFB patients versus the true one is provided in Table 14.

The classification, count and misclassification based on the Hierarchical Bayesian logistic regression analysis of the w6MW_d1 and DLCO with added random component of smoking status is shown in Table 14, and matches the results from Figure 6. It shows a high proportion of correct classification, with four subjects that are classified as patients when in fact they are supposed to be normal patients. The low level of DLCO significantly depict IPF patients, and therefore alter the lung functions.

The statistical models describe the useful first steps in making inferences about the data. While characterization of the IPF with most associated explanatory variables is best done in a designed study, the variables selected, even if non-significant at first, were now to be included in the suitable model. Overfitting may be a concern, mainly because of the sample size, although Bayesian method here can be considered as a comparative justifiable model [17,18].

## Conclusion

Although the methods are complementary, this research has shown that the MtDCN level is linked to the oxygen, smoking, and the Forced Vital Capacity and forced Expiratory Volume. Moreover, we have proposed a risk cohort model for patients with IPF. The model enables the practitioner to monitor the patients according to key characteristics such as the oxygen use and smoking status, and assess the evolution of the pa-

tients. We have proposed critical thresholds that may reflect disease severity/progression, and determine the clinical pathway of IPF with the diffusing Capacity of the lung for Carbone Dioxide (DLCO) and the MtDCN levels, under multiple regression and under Bayesian hierarchical models. In our predictive model that have suggested high predictive accuracy, surveillance can be raised when the patients' DLCO or MtDCN level reach a certain level. These findings show that the factors such as the six minute Walk Test (w6MW_d1), DLCO and smoking status are valuable indicators in predicting IPF progression, and without ignoring the FVC% at the 3rd and 12th months. All these predictors define a clinical subset of patients with and without IPF whose information can be used to tailor most appropriate surveillance cohort and medical treatments. We propose that any newly created medicine or treatments and their combinations should be developed while accounting for the influence of these factors. We have shown the statistical strength that can guide practitioners in the counselling and management of these patients.

Our study has some limitations as the number of patients is relatively small and our results should be validated in a larger independent population of IPF patients. Nonetheless, the development of our statistical models present a framework for how to incorporate factors that might influence the course and outcome of patients with this deadly, yet unpredictable disease.

## Acknowledgment

Financial support was provided by the Virginia 4VA grant at George Mason University and Old Dominion University, the Statistics in the Biomedical Acceleration of Idiopathic Pulmonary Fibrosis Research, and the Division of Intramural Research of the National Heart, Lung, and Blood Institute. We are grateful to all the participants and staff at both Inova Fairfax Hospital and the National Heart, Lung, and Blood Institute, and the National Institute of Health.

## References

1. Kirillov V, Siler JT, Ramadass M, Ge L, Davis J, et al. (2015) Sustained activation of toll-like receptor 9 induces an invasive phenotype in lung fibroblasts: Possible implications in idiopathic pulmonary fibrosis. Am J Pathol 185: 943-957.
2. Ryu C, Sun H, Gulati M, Herazo Maya JD, Chen Y, et al. (2017) Extracellular mitochondrial DNA is generated by fibroblasts and predicts death in idiopathic pulmonary fibrosis. Am J Respir Crit Care Med 196: 1571-1581.
3. Alder JK, Hanumanthu VS, Strong MA, DeZern AE, Stanley SE, et al. (2018) Diagnostic utility of telomere length testing in a hospital-based setting. Proc Natl Acad Sci 115: E2358-E2365.
4. Cronkhite JT, Xing C, Raghu G, Chin KM, Torres F, et al. (2008) Telomere shortening in familial and sporadic pulmonary fibrosis. Am J Respir Crit Care Med 178: 729-737.
5. Stuart BD, Lee JS, Kozlitina J, Noth I, Devine MS, et al. (2014) Effect of telomere length on survival in patients with idiopathic pulmonary fibrosis: An observational cohort study with independent validation. Lancet Respir Med 2: 557-565.
6. Carpagnano G, Lacedonia D, Cotugno G, Malerba M, Patricelli G, et al. (2017) Changes of mitochondria copy number in association with idiopathic pulmonary fibrosis. Clin Res Pulm 5: 1039.
7. Efron B (1994) Missing data, imputation, and the bootstrap. J Am Stat Assoc 89: 463-475.
8. Foote K, Reinhold J, Yu EPK, Figg NL, Finigan A, et al. (2018) Restoring mitochondrial DNA copy number preserves mitochondrial function and delays vascular aging in mice. Aging Cell 17: e12773.
9. Kim JH, Kim HK, Ko JH, Bang H, Lee DC (2013) The relationship between leukocyte mitochondrial DNA copy number and telomere length in community-dwelling elderly women. PLoS One 8: e67227.
10. Szklarczyk R, Nooteboom M, Osiewacz HD (2014) Control of mitochondrial integrity in ageing and disease. Philos Trans R Soc Lond B Biol Sci 369.
11. Raghu G, Richeldi L (2017) Current approaches to the management of idiopathic pulmonary fibrosis. Respir Med 129: 24-30.
12. McCullagh P, Nelder JA (1989) Generalized linear models. CRC press, 37.
13. Carlin BP, Kadane JB, Gelfand AE (1998) Approaches for optimal sequential decision analysis in clinical trials. Biometrics 54: 964-975.
14. Müller P, Berry DA, Grieve AP, Smith M, Krams M (2007) Simulation-based sequential bayesian design. JSPI 137: 3140-3150.
15. Rossell D, Müller P (2013) Sequential stopping for high-throughput experiments. Biostatistics 14: 75-86.
16. Goldstein H, Carpenter JR, Browne WJ (2014) Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. J R Stat Soc Ser A Stat Soc 177: 553-564.
17. Samara KD, Margaritopoulos G, Wells AU, Siafakas NM, Antoniou KM (2011) Smoking and pulmonary fibrosis: Novel insights. Pulm Med.
18. King TE, Tooze JA, Schwarz MI, Brown KR, Cherniack RM (2001) Predicting survival in idiopathic pulmonary fibrosis: Scoring system and survival model. Am J Respir Crit Care Med 164: 1171-1181.